Verification of Precipitation Reforecasts over the Southeast United States

Martin A. Baxter
Department of Earth and Atmospheric Sciences
Central Michigan University
Mt. Pleasant, Michigan

Gary M. Lackmann
Department of Marine, Earth, and Atmospheric Sciences
North Carolina State University
Raleigh, North Carolina

Kelly M. Mahoney
Cooperative Institute for Research in the Environmental Sciences
Univ. of Colorado Boulder
and
NOAA Earth System Research Lab, Physical Sciences Division
Boulder, Colorado

Thomas E. Workoff
NOAA/NCEP Weather Prediction Center
Systems Research Group, Inc.
College Park, Maryland

Thomas M. Hamill
NOAA Earth System Research Lab, Physical Sciences Division
Boulder, Colorado

Submitted to Weather and Forecasting
May 20, 2014

Corresponding author address: Martin A. Baxter, Dept. of Earth and Atmospheric Sciences, Central Michigan University, 314 Brooks Hall, Mt. Pleasant, MI 48859
E-mail: baxte1ma@cmich.edu

ABSTRACT

NOAA's second-generation reforecasts are approximately consistent with the operational version of the 2012 NOAA Global Ensemble Forecast System (GEFS). The reforecasts allow verification to be performed across a multi-decadal time period using a static model, in contrast to verifications performed using an ever-evolving operational modeling system. This contribution examines three commonly used verification metrics for reforecasts of precipitation over the Southeast United States: equitable threat score, bias, and ranked probability skill score. Analysis of the verification metrics highlights variation in the ability of the GEFS to predict precipitation across amount, season, forecast lead time, and location. Beyond day 5.5, there is little useful skill in quantitative precipitation forecasts (QPF) or probabilistic QPF. For lighter precipitation thresholds (e.g., 5 and 10 mm 24 h^{-1}), use of the ensemble mean adds about 10% to forecast skill over the deterministic control. QPFs have increased in accuracy from 1985 to 2013, mostly due to improvements in observations. Results of this investigation are a first step toward using the reforecast database to isolate weather regimes that the GEFS typically predicts well, and regimes that the GEFS typically predicts poorly.

1. Introduction

Attendant with the development of advanced numerical weather prediction (NWP) systems is the need to verify the capabilities of these systems. In particular, the quantitative precipitation forecast (QPF) is important to society and challenging for NWP. At what forecast lead time does NWP QPF skill effectively vanish? How much additional skill is provided by the use of ensemble forecasts, versus deterministic forecasts? Has QPF skill changed over time? The development of NOAA's second-generation reforecast database (Hamill et al. 2013) allows these questions to be addressed. Due to challenges in verifying QPF in areas where the precipitation climatology varies considerably (Hamill and Juras 2006), our focus is restricted to the southeastern United States (SEUS), where climatological precipitation characteristics are relatively homogeneous (Prat and Nelson 2013).

The SEUS receives precipitation associated with a variety of meteorological phenomena, including tropical cyclones, baroclinic waves, mesoscale convective systems, and localized diurnal convection (Moore et al. 2014). The juxtaposition of the Appalachian Mountains to the Atlantic Ocean and Gulf of Mexico creates an environment where QPFs are particularly challenging. While QPFs from NWP and forecasters have improved over the last 30 years (Novak et al. 2014), challenges remain. A static model run over multiple decades allows for verifications that span time scales longer than the periods that operational models remain unchanged. Long-term verification of reforecast data provides a sufficiently large sample size to allow the development of model climatology, which can then be compared with the climatology of the atmosphere.

NWP verification can inform forecasters of the strengths, limitations, and best applications of a modeling system. Ensembles attempt to provide a measure of confidence in a particular outcome, but if the model driving the ensemble system has difficulty in predicting a phenomenon, the ensemble spread can be misleading to a forecaster and therefore has reduced utility**.** Thus, it is helpful to document the accuracy of a modeling system's QPFs over long periods, and to provide this information to forecasters in a manner that allows for easy incorporation into the forecast process. As the value added by human forecasters over model QPF is diminishing (Novak et al. 2014), the need to leverage any sources of available information to improve upon model QPF becomes increasingly critical.

NOAA's second-generation reforecasts are approximately consistent with the operational 00 UTC cycle of the 2012 NOAA Global Ensemble Forecast System (GEFS). The 11-member reforecasts were created at ~0.5° grid spacing for week 1+ and ~0.75° grid spacing for week 2+ forecasts. Due to the change in grid spacing, we use only 0-7 day reforecasts. The reforecasts were verified using gridded precipitation data from the NOAA Climate Prediction Center Daily U.S. Unified Precipitation Dataset (Chen et al. 2008). This analysis is comprised of gauge observations interpolated onto a 0.25° grid. The 29 years (January 1985 – December 2013) of verification undertaken here precludes the use of a multi-sensor precipitation dataset. The reforecasts and precipitation data were interpolated to a common 0.5° x 0.5° grid over the SEUS (Fig. 6) using bilinear interpolation. As the period of observed precipitation is from 12 UTC – 12 UTC, and the reforecasts were initialized at 00 UTC, the lead times used in this study are 1.5 days (12-36 h) through 7.5 days (156-180 h). For in-depth explanation of verification metrics, the reader is directed to Wilks (2011, Ch. 8) and Jolliffe and Stephenson (2012, Chs. 2 and 3).

104 To enhance readability, all time-series of yearly quantities have been filtered with the 1-2-1
105 or "Hanning" filter (see Von Storch and Zwiers 1999).

106 **2. Verification of Deterministic Forecasts**
107
108 The equitable threat score (ETS) evaluates model ability to predict a two-category
109 event. Values range from -1/3 to 1, where 1 represents a perfect forecast and 0 or less
110 indicates unskilled forecasts. ETS is given by:

$$\frac{h - hc}{h + fa + m - hc}, \quad hc = \frac{(h + fa)(h + m)}{n} \tag{1}$$

111 Here, *h* are hits, *fa* are false alarms, *m* are misses, and *hc* represents hits correct by chance
112 (*n* is equal to *h* + *fa* + *m* + *cn*, where *cn* are correct negatives).

113 Time series of annual ETS depict an upward trend (Fig. 1). As the model and data
114 assimilation system used in the reforecasts remains static over the period, the upward
115 trend in ETS is mostly due to improvement of the initial conditions used by the model
116 through increased and higher quality observations. Alternatively, the upward trend may in
117 part be explained by changes in the quality of the verification dataset. The year-to-year
118 variability in ETS evidently results from variability in the model's ability to predict the
119 phenomena that lead to precipitation in those years. Also, ETS tends to increase with
120 fractional area coverage of the phenomenon (Hamill 1999; Moore et al. 2014). Model QPF
121 errors arise from both the quality of the model and observations, and the model
122 atmosphere's sensitive dependence to initial conditions (Zhang et al. 2006, after Lorenz
123 1996). ETS decreases with increasing lead time and with increasing threshold. Beyond day
124 5.5, average ETS values are below 0.1 (except for the 5 mm threshold), indicating little to
125 no skill. When the average ETS over 1985-1989 is compared with the average ETS over

2009-2013, an increase in ETS is seen for all lead times at all thresholds. When these 5-year average ETS values are averaged over the four thresholds in Fig. 1, the increase in ETS over the period for day 1.5 forecasts is 0.08. For day 7.5, the increase in ETS is 0.02. Thus, when considered across all four thresholds, day 1.5 ETS values have increased 3 times as much as day 7.5 ETS values over the nearly 30 year period. When averaged over all thresholds, the most recent day 3.5 reforecasts (ETS of .20) are equivalent in accuracy to the oldest day 1.5 reforecasts (ETS of .21). This increase in useful lead time is likely due to increases in the quantity and quality of observations, and illustrates the considerable impact these better observations have had on QPF over the SEUS.

Considerable seasonal variability in ETS is present for the 20 mm threshold (Fig. 2). The 20 mm threshold is chosen for further analysis, as it represents an "intermediate" amount of precipitation in the SEUS, and it occurs with sufficient frequency to allow meaningful analysis. Average ETS is highest in winter (0.19), followed by fall (0.16), spring (0.15), and summer (0.07), consistent with Fig. 3 of Hamill et al. (2013). This order in average ETS persists for days 1.5 through 5.5. At days 6.5 and 7.5, average ETS for winter, spring, and fall are essentially the same, within .01. Beyond day 5.5, ETS values for all seasons are below 0.1, indicating little to no skill. As with annual ETS, seasonal ETS exhibits an increase in 5-year average ETS from the beginning to the end of the period. The trend in summer day 1.5 ETS is minimal (.01), less than that of winter (0.13), spring (0.10), or fall (0.10), suggesting that improvements in observations have not lead to as much improvement in summer QPF compared with other seasons (e.g., Fritsch and Carbone 2004). Similar patterns are seen in ETS for other thresholds.

148 The bias score describes the ratio of the number of "yes" forecasts to the number of
149 "yes" observations, where h, fa, and m are as in (1):

$$B = \frac{h + fa}{h + m} \quad (2)$$

150 When bias exceeds one, the event is overforecast, and when bias is less than one, the event
151 is underforecast. A bias of one indicates the event was forecast the same number of times
152 as it was observed. Bias does not measure the correspondence between individual forecast-
153 observation pairs, and thus provides no information on model accuracy. Therefore, the
154 observed bias is independent of forecast lead time. At a 20 mm threshold, precipitation is
155 underforecast in all seasons (Fig. 3), as indicated by averages over all years and forecast
156 lead times: winter (0.94), spring (0.89), summer (0.62), and fall (0.72). The bias closer to
157 unity in the winter may result from the tendency for gauges to underestimate precipitation
158 that falls as snow (Rasmussen et al. 2012). Annual variability in the bias likely arises from
159 variation in the number of events forecast or observed from year-to-year. When 5-year
160 averages of bias (averaged over all lead times) from the beginning to the end of the period
161 are compared, only winter exhibits a trend in bias over 0.1, with a decrease in bias of 0.24.

162 **3. Verification of Ensemble Forecasts**
163
164 The ensemble mean versus control QPF was compared by calculating the ETS for
165 each season over the 29-year period for both forecasts, and then finding the percent
166 difference (Fig. 4). In general, the ensemble mean provides the most improvement over the
167 control for lower thresholds and shorter lead times. As thresholds increase, the ensemble
168 mean is more susceptible to smearing (i.e., multiple non-overlapping positions of the
169 precipitation maximum across ensemble members), leading to underestimation of the
170 magnitudes found in the control. This effect is most pronounced in summer, when rain

amounts are climatologically higher and precipitation is more localized. In all seasons but summer, and for all thresholds but 40 mm, the ensemble mean has an average of 6% higher ETS values versus the control through 4.5 days lead time. In all seasons but summer, the 5 and 10 mm thresholds have an average of 10% higher ETS values in the mean for all lead times through 7.5 days.

The ranked probability score (RPS) measures the difference between the cumulative distributions of forecasts and observations over a set of categories, as follows:

$$RPS = \sum_{k=1}^{K} \left(CDF_{fc,k} - CDF_{obs,k}\right)^2 \quad (3)$$

where k = 1 … K indicates the number of categories, and CDF refers to the cumulative distribution of either the forecast or observations. Categories were chosen to be similar to those used for probabilistic QPF at NOAA's Weather Prediction Center[1]. An RPS skill score is defined as $RPSS = 1 - \frac{\overline{RPS}}{\overline{RPS_{CL}}}$, where RPS_{CL} is the RPS computed using the cumulative climatological distribution to forecast the cumulative distribution of the observations. Overbars indicate that values are averaged over time and space. RPSS values less than 0 indicate that a PQPF from the ensemble is no better than using a climatological distribution as a probabilistic prediction. Seasonal values of RPS_{CL} were calculated at every grid point using cumulative climatological distributions for each season. Note that comparing the RPSS values applied here with other ensemble systems with differing numbers of members will not provide an even comparison, and additional calculations would be needed to appropriately make such comparisons (Richardson 2001).

[1] Categories are mutually exclusive and encompass all possibilities: ≥ 0 and < 1 mm, ≥ 1 and < 3 mm, ≥ 5 and < 10 mm, ≥ 10 and < 20 mm, ≥ 20 and < 25 mm, ≥ 25 and < 40 mm, ≥ 40 and < 50, ≥ 50 and < 65 mm, ≥65 and < 75 mm, and ≥ 75 mm.

As with ETS from the control run, when RPSS are averaged over all years and thresholds (Fig. 5), winter has the highest skill score (0.23), followed by fall (0.19), spring (0.14), and summer (-0.05). The negative summer value indicates that RPS values are worse than those that could be achieved by using climatology. In the summer, only day 1.5 RPS values exceed those from climatology. This result indicates that the resolution of the reforecast system is not designed to predict the scale of the phenomena dominant in the summer months in the SEUS; intrinsic predictability in the model is reduced in the presence of convection (Zhang et al. 2003). Systematic deficiencies in the GEFS also contribute to QPF error, such as the excessive semblance of ensemble members, resulting in over-confident probabilistic forecasts (Palmer 2012). Average RPSS values beyond day 5.5 are less than 0.1, indicating little to no skill. Seasonal RPSS experience an upward trend when 5-year average RPSS from the first five years are compared with those of the last 5 years. All forecast leads in every season exhibit this upward trend. This indicates that better observations have led to not only an increase in accuracy in deterministic QPF, but also in the ensemble's ability to provide a probabilistic QPF. The trend in day 1.5 RPSS is greatest for fall (.27), followed by winter (.20), spring (.12), and summer (.10). Interestingly, the summer exhibits a positive trend for PQPF as measured by RPSS, while no trend is observed for deterministic QPF as measured by ETS.

Spatial plots of seasonal RPSS (Fig. 6) qualitatively match the day 1.5 curves in Fig. 5, with winter and fall having the highest RPSS, and spring and summer the lowest RPSS. In the summer, PQPFs are no better than climatology when measured by RPS in parts of the domain. Outside of summer, areas in the interior of the domain feature the greatest

212 improvement over climatology. Longer lead times follow similar patterns, with the areas of
213 greatest improvement shrinking in size toward the center of the domain.

214 **4. Conclusions and Future Work**

215 In order to assess QPF skill as a function of lead time, and to quantify the benefit of
216 the ensemble mean over deterministic QPF, verification of the NOAA second-generation
217 reforecast dataset has been completed for the SEUS from 1985-2013. This long-term
218 verification provides forecasters with useful information on the capabilities of the current
219 generation of NOAA's GEFS. Both deterministic and probabilistic QPF exhibit long-term
220 increases, mostly due to the improvement of observations. For deterministic QPF, summer
221 does not feature an increase, while an increase in summer occurs for PQPF. For all seasons,
222 day 6.5 and 7.5 QPFs and PQPFs for the SEUS have *little to no skill* over random chance or
223 climatology, respectively. The use of the ensemble mean rather than the control offers
224 benefit on average, especially for lower thresholds and shorter lead times. The fact that
225 QPF and PQPF both decrease in accuracy in the same seasonal order suggests that
226 forecasters should not have the same confidence in PQPFs across seasons. For example, a
227 70% probability of 20 mm in summer is less likely to verify than a 70% probability in
228 winter, as both the model QPF and PQPF are worse in summer. Reforecasts can be used to
229 improve forecasts of precipitation through many techniques, such as the analog-based
230 technique of Hamill and Whitaker (2006) or logistic regression (Hamill et al. 2008).
231 Verification of these adjusted real-time forecasts has shown them to be superior to the
232 reforecasts themselves and the unadjusted real-time forecasts (Hamill et al. 2013).

233 Moore et al. (2014) produced a 10-year climatology of extreme precipitation events
234 in the SEUS and found that events associated with stronger dynamical forcing were more

accurately predicted by the reforecasts in terms of ETS, bias, and fractional area. This is corroborated by the present study of all precipitation events in the SEUS, as the summer events were poorly predicted and are more likely to be associated with weaker dynamical forcing. Future work will analyze patterns associated with the most and least accurate reforecasts of precipitation events in the SEUS. Greater understanding of how the modeled atmosphere differs from the real atmosphere will allow forecasters and researchers to identify situations where model guidance is likely to be poor. In addition, continuing analysis of the complex relationship between forecast precipitation, ensemble spread, and accuracy will help forecasters to better convert ensemble guidance into useful forecast confidence (Palmer 2012). The results of such analyses can help forecasters better allocate time and effort in improving model guidance, and would allow researchers to better allocate resources in improving observing and modeling systems.

Acknowledgements

We thank NOAA's Earth Systems Research Laboratory for support of this project, and access to their computer systems and data. The lead author's time was supported by Central Michigan University. This work was also supported by a NOAA grant to North Carolina State University. We thank NOAA/NCEP/CPC for their creation of the Unified Precipitation Dataset, and the U.S. Department of Energy for funding the creation of the Reforecast-2 dataset.

References

Chen, M., W. Shi, P. Xie, V. B. S. Silva, V E. Kousky, R. W. Higgins, and J. E. Janowiak, 2008: Assessing objective techniques for gauge-based analyses of global daily Precipitation. *J. Geophys. Res.: Atmos.*, **113 (D4)**, doi: http://dx.doi.org/10.1029/2007jd009132

Fritsch, M. J. and R. E. Carbone, 2004: Improving quantitative precipitation forecasts in the warm season: A USWRP research and development strategy. *Bull. Amer. Meteor. Soc.*, **85**, 955–965. doi: http://dx.doi.org/10.1175/BAMS-85-7-955

Hamill, T. M., G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau, Y. Zhu, W. Lapenta, 2013: NOAA's Second-Generation Global Medium-Range Ensemble Reforecast Dataset. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565. doi: http://dx.doi.org/10.1175/BAMS-D-12-00014.1

Hamill, T. M., R. Hagedorn, J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS Ensemble Reforecasts. Part II: Precipitation. *Mon. Wea. Rev.*, **136**, 2620–2632. doi: http://dx.doi.org/10.1175/2007MWR2411.1

Hamill, T. M., and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.*, **134**, 3209-3229. http://dx.doi.org/10.1175/mwr3237.1

281

282 Hamill, T. M., and J. Juras, 2006: Measuring forecast skill: is it real skill or is it the varying
283 climatology? *Q. J. R. Meteorol. Soc.,* **132**, 2905–2923.
284 doi: http://dx.doi.org/10.1256/qj.06.25

285

286 Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea.*
287 *Forecasting*, **14**, 155-167. http://dx.doi.org/10.1175/1520-
288 0434(1999)014<0155:HTFENP>2.0.CO;2

289

290 Jolliffe, I.T., and D. B. Stephenson, eds, 2012: *Forecast verification: A practitioner's guide in*
291 *atmospheric science*. John Wiley & Sons, 274 pp.
292 doi: http://dx.doi.org/10.1002/9781119960003

293

294 Lorenz, E. N., 1996: Predictability- A problem partly solved. *Proc. Seminar on Predictability*,
295 Vol. I, Reading, United Kingdom, ECMWF, 1-19.
296 doi: http://dx.doi.org/10.1017/cbo9780511617652.004

297

298 Moore B. J., K. M. Mahoney, E. M. Sukovich, R. Cifelli, and T. M. Hamill, 2014: Climatology
299 and environmental characteristics of extreme precipitation events in the
300 Southeastern United States. *Mon. Wea. Rev.*, accepted pending minor revisions.

301

302 Novak, D. R., C. Bailey, K. Brill, P. Burke, W. Hogsett, R. Rausch, M. Schichtel, 2014:
303 Precipitation and temperature forecast performance at the Weather Prediction

Center *Wea. Forecasting*, in press. doi: http://dx.doi.org/10.1175/WAF-D-13-00066.1

Palmer, T. N., 2012: Towards the probabilistic Earth-system simulator: A vision for the future of climate and weather prediction. *Q. J. R. Meteorol. Soc.,* **138**, 841–861. http://dx.doi.org/10.1002/qj.1923

Prat, O. P., and B. R. Nelson, 2014: Characteristics of annual, seasonal, and diurnal precipitation in the southeastern United States derived from long-term remotely sensed data. *Atmos. Res.,* **144**, 4–20. doi: http://dx.doi.org/10.1016/j.atmosres.2013.07.022

Rasmussen, R. M., and coauthors, 2012: How well are we measuring snow: The NOAA/FAA/NCAR Winter Precipitation Test Bed. *Bull. Amer. Meteor. Soc.,* **93,** 811–829. doi: http://dx.doi.org/10.1175/bams-d-11-00052.1

Richardson, D. S., 2001: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Q. J. R. Meteorol. Soc.,* **127**, 2473–2489. doi: http://dx.doi.org/10.1002/qj.49712757715

Von Storch, H., and F. W. Zwiers, 1999: *Statistical Analysis in Climate Research*. Cambridge University Press, 484 pp. http://dx.doi.org/10.1017/CBO9780511612336

Wilks, D. S., 2011: *Statistical methods in the atmospheric sciences*. Vol. 100. Academic Press, 676 pp.

Zhang, F., C. Snyder, and R. Rotunno, 2003: Effects of moist convection on mesoscale predictability. *J. Atmos. Sci.*, **60**, 1173-1185. doi: http://dx.doi.org/10.1175/1520-0469(2003)060<1173:eomcom>2.0.co;2

Zhang, F., A. M. Odins, and J. W. Nielsen-Gammon, 2006: Mesoscale predictability of an extreme warm-season precipitation event. *Wea. Forecasting*, **21**, 146-166. doi: http://dx.doi.org/10.1175/waf909.1

List of Figures

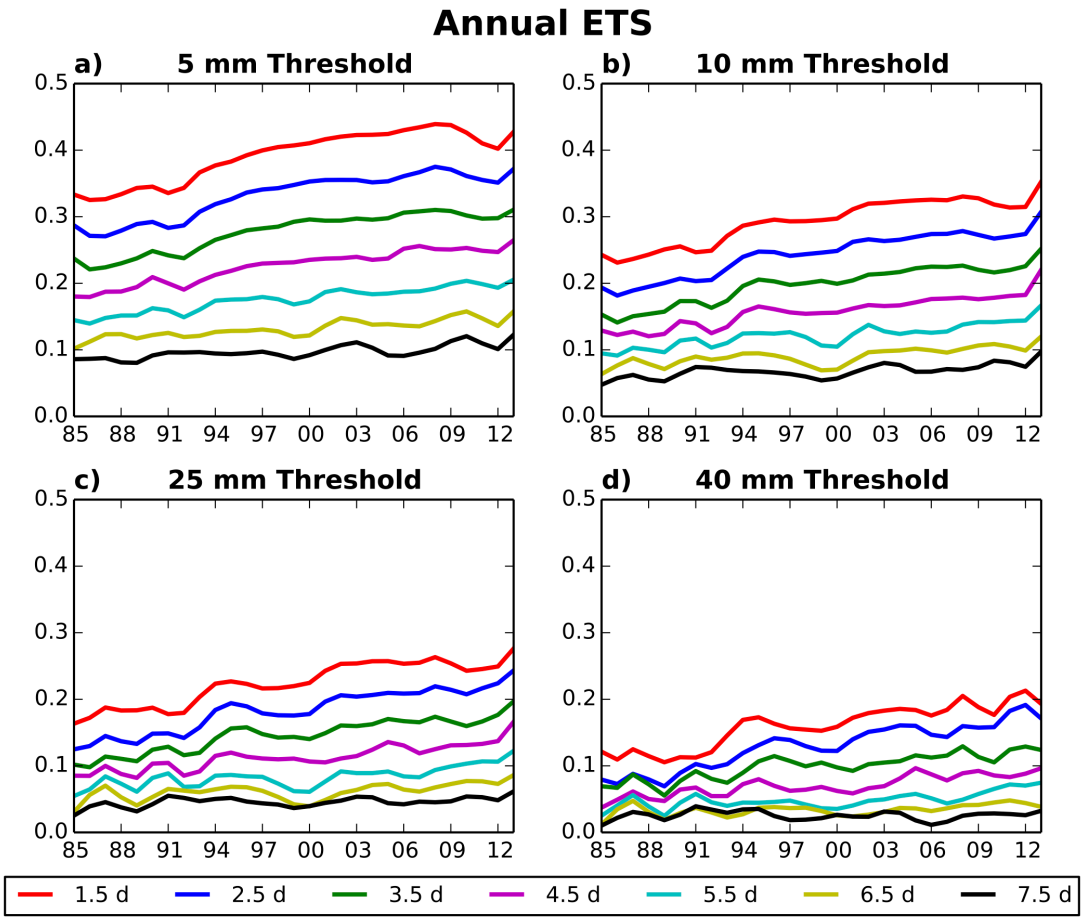

Figure 1: Annual equitable threat score from 1985-2013 for a) 5 mm threshold, b) 10 mm threshold, c) 25 mm threshold, and d) 40 mm threshold. Lines on graph are forecast lead times as indicated in legend. Domain used for calculation can be seen in Fig. 6.

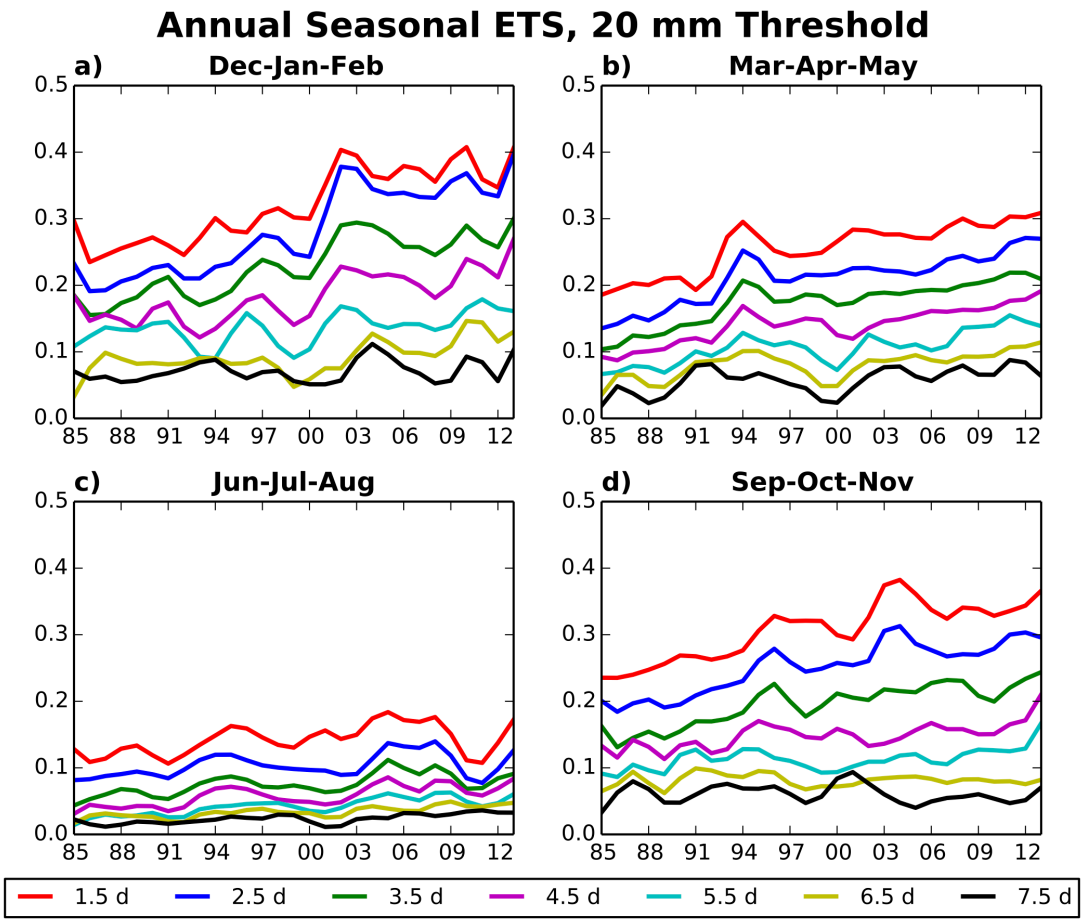

Figure 2: Annual equitable threat score for 20 mm threshold from 1985-2013 for a) winter, b) spring, c) summer, and d) fall. Lines on graph are forecast lead times as indicated in legend.

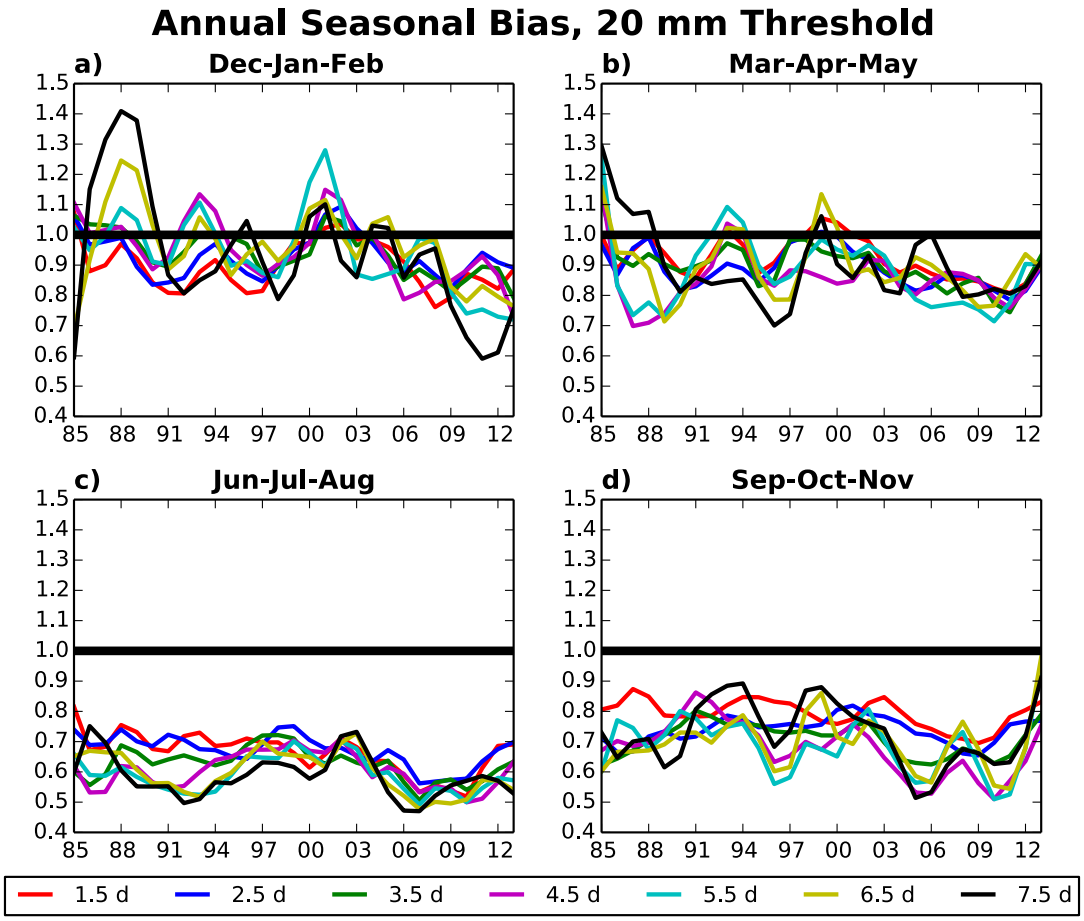

Figure 3: Annual bias for 20 mm threshold from 1985-2013 for a) winter, b) spring, c) summer, and d) fall. Lines on graph are forecast lead times as indicated in legend.

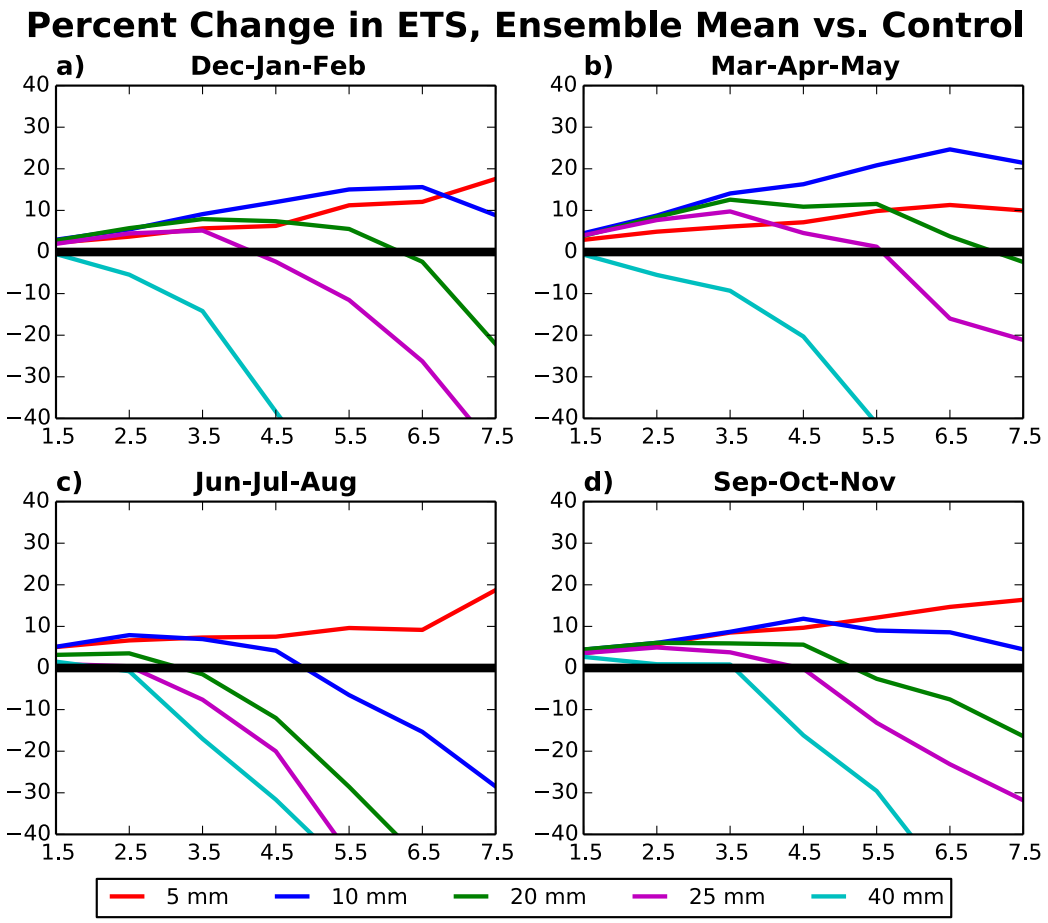

Figure 4: Percentage change in equitable threat score with forecast lead time when ensemble mean is compared with control for a) winter, b) spring, c) summer, and d) fall over the 1985-2013 period. Lines on graph are precipitation thresholds as indicated in legend.

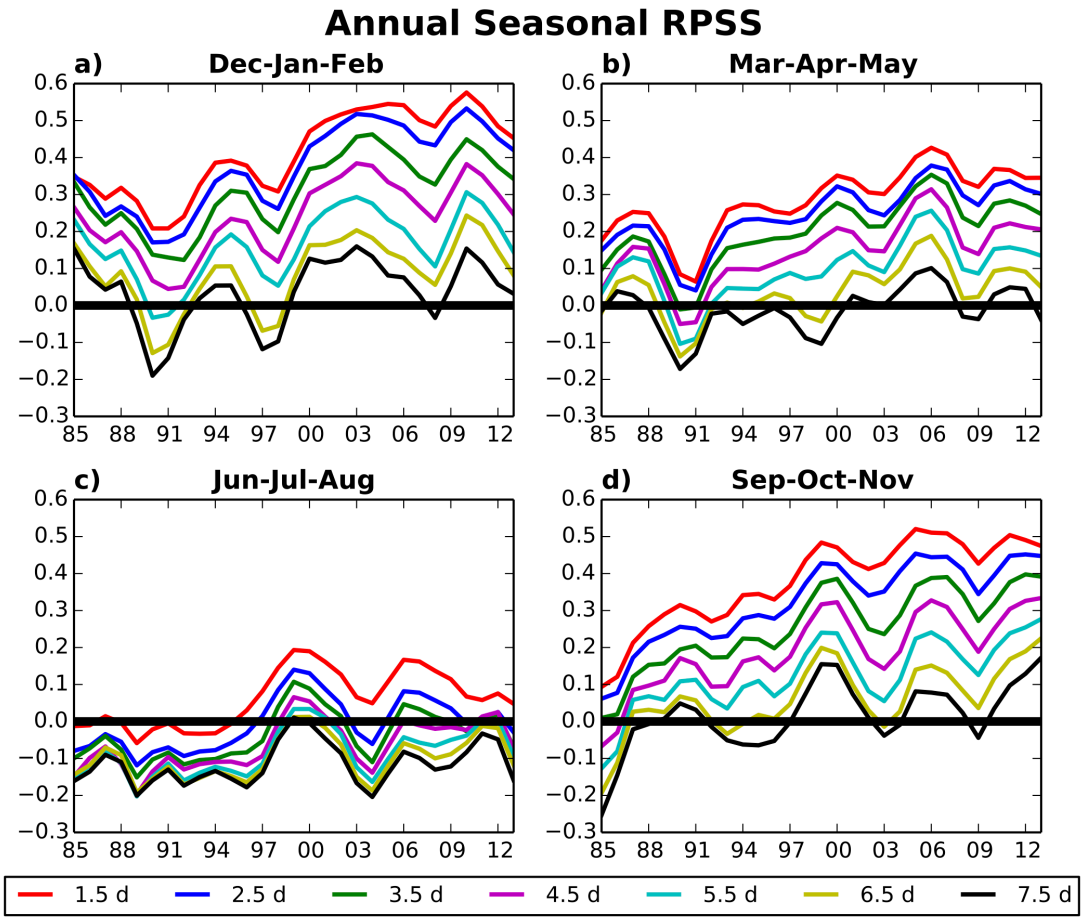

Figure 5: Annual ranked probability skill score from 1985-2013 for a) winter, b) spring, c) summer, and d) fall. Lines on graph are forecast lead times as indicated in legend.

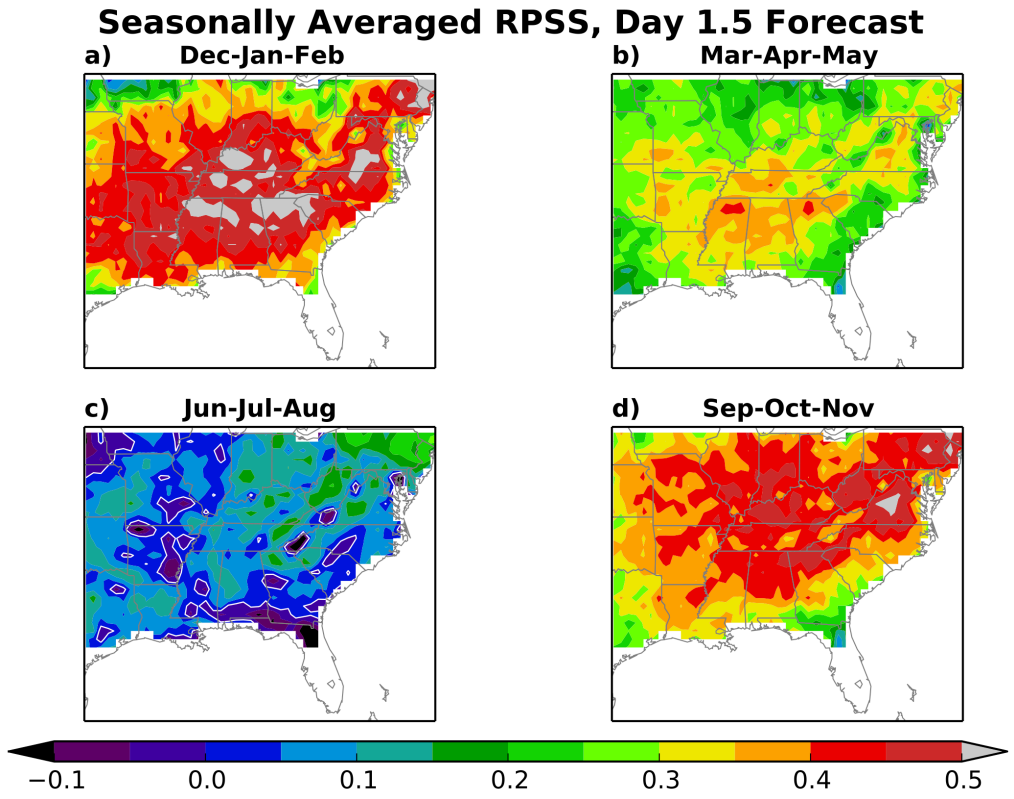

Figure 6: Day 1.5 ranked probability skill score from 1985-2013 for a) winter, b) spring, c) summer, and d) fall. Values less than zero contoured in white.